Early Machine Translation in France

Maurice Gross

1

Early machine translation in France

Maurice Gross

Université Paris 7, Laboratoire d'Automatique Documentaire et Linguistique

In 1960, I took my first job at the Computer Center of the Laboratoire Central de l'Armement (LCA), in Arcueil, near Paris. LCA belongs to the French Army. The Center was headed by Armand Sestier, a high-ranking engineer of the Armament Corps. At that time, the Computer Center was mainly handling computations of missile trajectories.

A. Sestier had been informed through NATO channels that a research and development programme on Mechanical Translation was launched by military agencies in the US and that NATO members could join. The French contribution took shape through an agreement between the Armament Corps and the Centre National de la Recherche Scientifique (CNRS), the national academic research agency. A division of the Computer Center of LCA became dedicated to MT and called Centre d'Étude pour la Traduction Automatique (CETA). Very soon, another center was created by the CNRS at the University of Grenoble and headed by Professor Bernard Vauquois, a computer scientist, specialist of programming languages. The centers were named CETAP and CETAG, P for Paris and G for Grenoble[1]. Paris specialized in Russian-French translation, whereas Grenoble was mostly devoted to the pair German-French. Although contacts between their directors were frequent, both laboratories operated quite independently. The computer equipment at CETAP was limited, its modern part was an IBM 650 model with an important novelty: a FORTRAN compiler, but programming it in assembler language was the rule. Contacts with linguists were mainly with specialists of Russian. Émile Delavenay, an official at UNESCO, in Paris, had some general linguistics background and had written an early essay on MT, Michael Corbe, another official at UNESCO had also an interest in MT, Both were often consulted by A. Sestier. Linguists with a good knowledge of Russian were recruited and started working on the morphology of Russian. Programmers were available, they learned to cooperate with linguists. As a beginner, I was in the same situation, I had volunteered to enter the CETA group, without any knowledge of linguistics, my main motivation was the bizarre nature of the activity and some taste for literature and language. I started reading and working for about one year, mostly on the analysis of Russian. Among my readings was Noam Chomsky's *Syntactic Structure*, which I could not well understand, but which attracted me as being relevant to the questions that MT raised. An opportunity came: I was offered a *UNESCO* fellowship to study MT in the United States. We had contacts with Anthony G. Oettinger of *Harvard University* and Victor Yngve of *MIT*. I left for Cambridge, Mass. in Fall 1961 for one year. I registered as a graduate student at *Harvard University*, and at the same time became a guest in V. Yngve's group. I followed different courses and seminars, in logic (recursive function theory, based on Martin Davis' remarkable book: *Computability and Unsolvability*), in computer science (at MIT, I was able to write (mini-)parsers in LISP and in COMIT, a language well-adapted to linguistic descriptions; COMIT was to be replaced by

[1] Later to become GETA.

SNOBOL and PROLOG) and in linguistics (phonology with Morris Halle and, most important, the fascinating courses in syntax by Noam Chomsky).

Everything I learned appeared to be necessary to solve some of the problems of MT, but it became quickly clear that all this knowledge was quite insufficient. Besides, open criticisms of MT were voiced by N. Chomsky and Y. Bar-Hillel on a theoretical basis, and even V. Yngve saw the aim of MT as very remote. Back in France, in Summer 1962, I presented a rather negative picture of the field, which, combined with a keen perception by CETA members of the difficulties of the task, convinced A. Sestier, the head of our group, to abandon MT altogether. Early in 1963, CETA was closed, and I was offered an assistantship at the CNRS, in the Institut Blaise Pascal, an academic research center in Computer Science.

I was mainly working on formal grammars, Marcel-Paul Schützenberger was my adviser. At the same time, I kept studying French syntax, more specifically, the problem of sentential complements in French. M.-P. Schützenberger introduced me to Zellig S. Harris, who invited me at the University of Pennsylvania, for the year 1964-5. Encouraged by both, I started a program of systematic description of French grammar. I spent the summer 1965 at MIT, talking to Yuki Kuroda, Ted Lightner, Jim McCawley, Paul Postal, John R. Ross and to Peter S. Rosenbaum[2] who was writing a dissertation on this subject in English.

When the ALPAC report was published (Pierce 1966), I was deeply convinced that MT made no sense in the absence of detailed and formalized language descriptions. Although the report did not particularly discuss this point, it had the advantage of reminding the members of the field that the aim of TA was entirely practical: a computer programme had to produce translations of a quality comparable to those of a human translator, or cheaper if the quality was lower. For theoreticians, computer scientists and some linguists, this was brutal awakening. Actually, the matter could have been discussed seriously much earlier. After all, MT development was an engineering task, combining computer programming and linguistics, two fields that had an autonomous life and from which MT developers had to start. For computer specialists, several new tasks were clear: new programming tools should be helpful, as well as new algorithmic tools (e.g. a hashing method proposed by Ted Ziehe at the *Rand Corporation*) and the size of the data (dictionaries, either monolingual or bilingual) meant that new types of memories would be necessary (e.g. the optical disk developed by Gilbert King at *IBM*). For linguists, several subfields of linguistics were involved: the synchronic description of each language, namely, its morphology and lexicon, its syntax and possibly its semantics. Besides, comparisons between pairs or families of languages were going to be used in the translation or transfer process. Individual descriptions of many languages existed and even though they had only been devised to teach first and second languages, common thought was that they were sufficient, up to some formalization step that required the choice of a formal model, adequate from the computational point of view. Actually, the choice was between variants of phrase structure grammars, although transformational linguists had criticized such models as linguistically inadequate. No inventory of the needs and of the resources had been made seriously. Quite logically, the simplest problems were first dealt with, for example, Russian morphology attracted many

[2] And many others who were finishing their dissertation and were going to become major generative linguists.

specialists. Different manuals that described conjugations and inflections were available, they had to be completed and formalized beyond the tables that had been compiled by traditional grammarians for students of Russian. But it should have been obvious that ambiguity was the major problem, and that only a detailed exploration of the contexts of ambiguous words could bring a solution. Systematic analysis of strings of words is syntax: grammar textbooks were available, they contained rules and exceptions, but when it came to confront them with texts, tremendous gaps appeared. For example, in many languages, the shapes of verb complements vary enormously, prepositions are unpredictable as every learner of a second language knows. No systematic list of the prepositions that go with a given verb had been compiled for any language, although this information seems valuable for language teaching. Obviously, this kind of data was going to be crucial to parse sentences automatically. In the same way, comparative studies of pairs of languages are rare, most of them deal with evolution phenomena and are limited to shifts in sound and in meaning, some etymological and philological studies are oriented to language teaching, none is of any use to MT.

As a matter of fact, a philosophical parameter has obscured the issue. Some specialists advocate the construction of specific MT programmes for ordered pairs of languages (i.e. source and target languages), hence taking advantages of similarities between some language pairs. Others think such an approach is too costly and propose to construct intermediary abstract languages into which a given source language would be translated and from which a translation would be generated into the target language (see fn 6). MT research never went far enough to demonstrate whether such a pivot language could exist, in fact, there are reasons to believe that it does not[3]. Another important resource is the dictionary: dictionaries do not really belong to the academic world of linguistics. They are commercial products, built to orders for a company and aimed at the general public. Users are human beings that can always supply a lot of information[4] to the definitions and to the limited examples that are provided. Lexicographers have established traditions about the nature and the amount of information put in regard of each dictionary entry. The problem is that lexicographic traditions vary from author to author, from publisher to publisher, hence they is no reason to believe that words are described in a standardized way, a first step toward the formalized forms needed by a computer programme to use the content of an entry. Another negative feature of most bilingual dictionaries was the lack of contexts for the word entries. In the last ten years, the situation has improved, and for the pair French-English, at least two excellent dictionaries are available to-day which include large number of phrases, that is, compound words or collocations that examplifiy specific meanings of simple words. It is worth mentioning that in in the same period, large dictionaries of idioms were developed in Eastern Europe on a regular basis, a type of work perhaps favored by Marxist thinking and/or the needs and weight of professional translators.

Another crucial component of MT is ideological. MT was initiated by the political and

[3] For a review of such artificial languages, see Couturat & Leau 1903. And since 1903, the activity has been going on smoothly.

[4] For example, the user of a dictionary who looks up a word has to perform a lemmatizarion of the word as it is found in a text, namely, a morphological analysis which can be quite complex to reproduce by computer.

military establishment around the 60s, in order to fill an intelligence gap. The needs were enormous, when calculated in numbers of pages to be translated. The idea that computer techonology was going solve the problem is not a new myth. The idea that machines could replace human beings in some intellectual activities has occupied the mind of many, and today, Artificial Intelligence is a concept taken for granted by the general public. I suppose that some high ranking officials did bet on an idea which, in a way, minimizes the risks: the amount of money to be put in the enterprise was so small within the budgets that research organizations were handling, that a failure would not hurt the least.[5] Such a way of reasoning avoids a full review of the scientific background necessary to resolve the problem. I suppose, that the Star Wars programme was decided in a similar way, the needs for new weapons was decided politically, irrespective of the availability of the physics underlying the guns that would shoot death rays (i.e. X-ray lasers). Actually, ideologically-driven MT enterprises did surface again. After the decline that followed the publication of the ALPAC report, mainly in the U.S., MT went on in Europe. In the 70s, when Canada went bilingual, French became an official language a par with English, as a consequence, all official documents had to be translated from English to French. Again, it was immediately apparent that the needs for translations topped the possibilities of human translators, again MT was invoked to solve the problem, again, failure led to abandon the few projects that had been set. And in the 80s, The *European Community* faced the same political problem: all languages of the *Community* are legally equal, hence all official documents must be available in the 11 languages of the Community[6]. Again, it became obvious at some point that the needs for translations overwhelmed the possibilities of human translators, again MT was invoked to solve the problem. A rather large research and development programme called EUROTRA was set up, it failed completely, and was abandoned a few years ago.

MT has not disappeared and it won't, it still fascinates people and officials. It is also interesting to notice that products do exist and sell. The best known programme is perhaps SYSTRAN°, one of the first commercial programmes. Its development was sponsored by the US Air Force and used for browsing through newspapers such as the Soviet *Pravda*. It was later developed in a commercial product. One version is used by the *European Community* as a way to get an approximation of some translations, which are then improved by human translators. Savings in the global translation process are thus made, the figure 20 % has been mentioned at occasions. SYSTRAN° is available in France to the general public through the MINITEL network, it is used in INTERNET browsers (ALTAVISTA). Many other programmes are being proposed, as services or as products that can be hooked on a word processor. The usefulness of the devices is entirely in the eye of the user, as

---

[5] The MT programme METAL, developed at SIEMENS, in Germany, was of a different nature, since the company had large needs in translation of technical documentation.

[6] In order to do some savings, the EUROTRA project had adopted the use of an intermediary language. The savings are argued for as follows: for 11 languages, one needs 11x10 = 110 translation programmes (remember that a procedure that goes from English to French is different from the procedure that goes from French to English). With an intermediary language IL, one has to translate from each of the 11 languages to the language IL, and then, from IL to each of the 11 languages, which requires the construction of only 22 MT procedures. A saving of 88 programmes!

it is extremely difficult to devise evaluation procedures for the quality of translations (Lehrberger & Bourbeau 1988). After all, in some circumstances, I might prefer a word to word translation from Chinese to French rather than no translation at all.

End of 1966, I became in charge of a research laboratory of the CNRS: the *Section d'Automatique Documentaire*, previously headed by Jean-Claude Gardin, an archeologist who developed methods of Information Retrieval of a general type. The field of IR appeared to me as presenting features similar to those of MT. I was convinced that the quality of full-text search[7] would benefit from detailed linguistic descriptions, and with the agreement of research officials of the *CNRS*, the laboratory was renamed *Laboratore d'Automatique Documentaire et Linguistique* and deeply involved in the study and formalization of French. A full research programme of systematic description was defined and became fully operational within a year. Linguists were hired and trained to produced, among other descriptions, systematic tables of French verbal constructions (Boons, Guillet & Leclère 1976,1988).

It is interesting that our main thesis which recommends basing future MT programmes on appropriate linguistic resources is far from being accepted. Many products on sale are almost empty shells that the user has to fill with the vocabulary he chooses to use. Grammars are never proposed on a similar basis. To me, the reason is clear: the language data currently available are not **cumulative**. Lexicographers did accumulate list of words and definitions, however, every dictionary is still the work of a single author; when an author disappears, often his dictionary disappears. A given author may be able to maintain some coherence throughout a dictionary, even when if he directs a team of lexicographers, but it is hard to conceive how two dictionaries, put out by two different publishers, could be merged into one single more complete dictionary. In particular, how could be removed the information common to both works? How could be merged the information that differ in both works? Lack of rigorous standards forbids merging operations. It is not unconceivable that standards for the presentation of a dictionary article be proposed, John Sinclair's has systematized the content of the COBUILD dictionary in this direction, but one cannot see competing publishers adopting such a common format. The situation is even worse for grammars, it would not come to anybody's mind to merge two existing grammars in order to produce a larger one. However, the very discussions of linguistic models of grammars might suggest that descriptive standards for syntactic phenomena could emerge. Formal models of grammars have been discussed as early as the years 1950s, but the discussions were never followed by a systematic programme of description for languages that qualify for MT. Typical of this situation is the state of generative grammar where the production is limited to polemic contributions aiming at proving the superiority of a variant of model over some other one, on the basis of a handful of examples, preferably drawn from languages almost extinct or accessible only at the end of a trek in the Andean Cordillera (M. Gross 1978).

We consider that authors such as L. Bloomfield, N. Chomsky and Z.S. Harris have provided the methodology for building cumulative lexicons and grammars. Namely,

[7] At that time, full text meant titles of books or articles, at best abstracts. It is only recently that full-text electronic libraries are considered to be set up.

that given a set of principles, different linguists in different teams could operate descriptions of words and structures that could be merged, for example in one same databank. There is a price to pay: descriptions should be limited to **reproducible** phenomena, which is precisely what the above mentioned authors have attempted to clarify. In particular, no semantics is within reach today, but most morphological phenomena can be handled, especially for languages whose spelling is fixed. Also, a large number of syntactic constraints can be described: all those that rely on the **reproducible intuition of acceptability**. Models that are quite neutral from the theoretical point view can be offered, finite-state automata for example (Roche & Schabes 1997). Data presented in such a format can be retranslated into any other formalism, more specific or more powerful, according to anyone's tastes. Experienced has shown that grammar rules described in this way can be accumulated with a precision such that they can be directly included into an automatic parser (M. Gross 1989, 1997).

Based on the principle that only reproducible phenomena can be accumulated (i.e. in a scientific way, namely, coherently and systematically), we have defined a research programme called RELEX adopted by a network of research laboratories in Europe. Each team describes its native language, among the advanced descriptions are: Bulgarian, English, French, German, Greek, Italian, Polish, Portuguese, Serbo-Croatian and Spanish[8]. For these languages, dictionaries of about 100,000 entries have been constructed, inflected forms are automatically derived from the entries and the morphological codes assigned to them. Dictionaries of compound words have also been assembled for all parts of speech, they are inflected in order to be matched with text units. We provide various samples of these dictionaries in the annex. Using the grammatical categories of the dictionaries, grammars are being constructed for all these languages, showing that several authors can independently write local grammars that can be merged (automatically) and that provide coherent parsers. Again, in the annex, we present samples of such grammars.

All these data are currently in use in a corpus parser called INTEX (M. Silberztein 1993). Procedures of representation common to all the mentioned languages are imposed by INTEX, they can be seen as a point of departure for further standardization of linguistic representations.

[8] A Korean and a Thaï dictionary are under way.

Annex

CAPTIONS

1) For the table with '+' and '-' marks:

A sample of the lexicon-grammar of French:

2) For the other page:

A sample of electronic dictionaries of simple and compound words. Codes correspond to basic grammatical categories: part of speech and morphological classes which are used to inflect automatically the entries, essentially according to gender, number, tense, mood.

References

Boons, Jean-Paul ; Alain, Guillet ; Christian, Leclère 1976. *La structure des phrases simples en français. Vol I : Constructions intransitives*, Geneva: Droz, 377p.

Couturat, L. & L. Leau, 1903. *Histoire de la langue universelle*, Paris: Hachette, 574 p.

Gross, Maurice 1975. *Méthodes en syntaxe*, Paris: Hermann, 412p.

Gross, Maurice 1978. On the Failure of Generative Grammar, *Language*, Vol. 55, N° 4, Baltimore: The Waverly Press.

Gross, Maurice 1989. The use of finite automata in the lexical representation of natural language, In *Electronic Dictionaries and Automata in Computational Linguistics*, Berlin-New York: Springer Verlag, pp.18-34.

Gross, Maurice 1997. The Construction of Local Grammars, In Roche & Schabès, eds. 1997. *Finite State Language Processing*, Cambridge, Mass.: The MIT Press, pp. 329-352.

Guillet, Alain ; Christian Leclère 1988. *La structure des phrases simples en français. Vol II : Verbes à complément direct et complément locatif*, Geneva: Droz, 445 p.

INTEX, Presentation and demos available at **http://www.ladl.jussieu.fr**.

Lehrberger John & Laurent Bourbeau, 1988, *Machine translation, Linguistic characteristics of MT Systems and general methodology of evaluation*, Lingvisticae Investigationes Supplementa, 15, Amsterdam-Philadelphia: John Benjamins BV.

Pierce, John R. ed. 1966, *Language and machines, Computer in translation and linguistics*, Washingon D.C.: National Academy of Sciences, 124 p.

Roche, Emmanuel & Yves Schabès, eds. 1997. *Finite State Language Processing, Cambridge*, Mass.: The MIT Press.

Silberztein, Max 1993. *Dictionnaires électroniques et analyse automatique de textes: le système INTEX*. Paris: Masson, 233 p.